การทำเหมืองข้อมูลเพื่อเพิ่มประสิทธิภาพให้องค์กร

Data mining to optimize the organization

ชัยศิริ สนิทพลกลาง

สาขาวิชาวิทยาการคอมพิวเตอร์ มหาวิทยาลัยราชภัฏจันทรเกษม ; chaisanit95@gmail.com

บทคัดย่อ

ปัจจุบันการใช้ประโยชน์จากข้อมูลต่างๆ เพื่อวัตถุประสงค์ที่เฉพาะเจาะจงนั้นมีความต้องการมากขึ้น โดยปริมาณข้อมูลที่มีอย่างมหาศาลนั้นประกอบไปด้วยข้อมูลที่สำคัญเป็นประโยชน์ต่อการนำมาใช้ และอีกทั้งยังมีข้อมูลที่ไม่มีความจำเป็นเช่นกันที่ปะปนกันอยู่ซึ่งทำให้มีความซับซ้อนและปัญหาในการเลือกนำมาใช้เพื่อเพิ่มศักยภาพ และขีดความสามารถในการแข่งขันแก่องค์กรนั้นๆ ปัจจุบันวิทยาการข้อมูลเป็นศาสตร์หนึ่งที่สามารถช่วยแก้ปัญหาและจัดการข้อมูลมหาศาลเหล่านั้น ซึ่งวิทยาการข้อมูล เป็นการผนวกศาสตร์ อาทิ สถิติ คณิตศาสตร์ และวิทยาการคอมพิวเตอร์เข้าด้วยกัน เพื่อสร้างเครื่องมือและวิธีการในการคัดแยก แจกแจงข้อมูลต่างๆ ซึ่งจะเป็นประโยชน์สูงสุดในการนำข้อมูลเหล่านั้นไปใช้ การสกัดข้อมูลสำคัญและเป็นประโยชน์ จากข้อมูลขนาดใหญ่ต่างๆ เป็นหนึ่งในวิธีการของการนำเอาวิทยาการข้อมูลมาใช้เพื่อดึงเอาข้อมูลที่สำคัญและเฉพาะเจาะจงไปใช้ในการวิเคราะห์ผลกระทบต่างๆขององค์กร เช่น ผลกระทบเชิงธุรกิจ การค้าขาย ผลกำไร เป็นต้น ดังนั้นบทความวิชาการนี้ได้รวบรวมและอธิบายเกี่ยวกับความเข้าใจและวิธีการสกัดข้อมูลสำคัญจากข้อมูลขนาดใหญ่ขององค์กร เพื่อนำมาใช้ให้เกิดประโยชน์สูงสุดต่อองค์กร

คำสำคัญ : เหมืองข้อมูล, ข้อมูลขนาดใหญ่, วิทยาการข้อมูล, Apirori algorithm

Abstract

Nowadays, the utilization of data for a specific purpose has been exponentially increased, as known that there is an enormous database (big data), which has been typically composed a mixture of either necessary data or unnecessary data that poses a difficulty to utilize only the specific data for an organization's specific purposes. The selection of necessary data from big data aims to increase the performance and the competitiveness of organizations. However, as mentioned the mixture of necessary and unnecessary data make the difficulty to classify and individualize data separately. A data science therefore, plays an important role to establish either tools or methods that would facilitate the separation of a mixture of data by a combination of statistic, mathematics and computer science together. Extraction method of necessary data is one of foundational methodology of data science, which will help to acquire only necessary data from a big data for a specific purpose. This is used to extract and analyze the factors that affect the organization performances for instance, the effect on business-marketing and the organization benefits. This review article has an objective to describe and overview an overall of an extraction method of necessary data in order to maximize the utilization of necessary data.

Keywords: Data Mining, Big Data, Data Science, Apirori algorithm

1. บทนำ

ความก้าวหน้าของเทคโนโลยีที่ก้าวไปอย่างรวดเร็วในปัจจุบันไม่ว่าจะเป็นเทคโนโลยีทางด้านการศึกษา ธุรกิจ สิ่งแวดล้อม ยานยนต์ การคมนาคม และด้านอื่นๆ อีกมากมาย รวมถึงการติดต่อสื่อสารเพื่อแลกเปลี่ยนข้อมูลล้วนเป็นการเพิ่มปริมาณข้อมูลให้มีจำนวนมหาศาล หรือที่เรียกกันว่า “บิ๊กดาต้า”(Big Data) โดยข้อมูลเหล่านี้อาจมีลักษณะทางกายภาพแบบมีโครงสร้าง แบบไม่มีโครงสร้าง และแบบกึ่งโครงสร้าง ขึ้นอยู่กับบริบทของเทคโนโลยีนั้นๆ จะนำเสนอข้อมูลในรูปแบบไหน ซึ่งข้อมูลขนาดใหญ่ก็จะปะปนไปด้วยข้อมูลที่เป็นประโยชน์และข้อมูลที่ไม่มีความจำเป็นต่อการนำไปใช้ประโยชน์ โดยข้อมูลที่เป็นประโยชน์เราสามารถนำมาใช้ให้เกิดประโยชน์ต่อองค์กรได้ เพื่อเป็นการเพิ่มประสิทธิภาพการทำงาน จากคำกล่าวที่ว่า “Information is power ”[1] คือ อำนาจที่เกิดจากการมีข้อมูลสารสนเทศจะทำให้องค์กรเราได้เปรียบเหนือคู่แข่ง

ดังนั้นการค้นหาข้อมูลที่เป็นประโยชน์จากข้อมูลขนาดใหญ่นั้นจึงมีความจำเป็นสำหรับองค์กร ซึ่งไม่ใช่เรื่องง่ายนักที่จะทำได้ แต่ในปัจจุบันได้เกิดศาสตร์ที่เรียกตัวเองว่า “วิทยาการข้อมูล” เกิดขึ้นมาโดยการผนวกศาสตร์ ด้าน สถิติ คณิตศาสตร์ และวิทยาการคอมพิวเตอร์ เข้ากับองค์ความรู้ที่เป็นความเชี่ยวชาญเฉพาะด้านต่างๆ เพื่อเข้ามาช่วยสกัดหาข้อมูลที่เป็นประโยชน์จากข้อมูลขนาดใหญ่ โดยใช้หลักสถิติและ คณิตศาสตร์ เข้ามาช่วยสร้างความน่าเชื่อมั่นและความถูกต้องกับข้อมูลที่เป็นประโยชน์ ส่วนวิทยาการคอมพิวเตอร์จะนำมาช่วยในการจัดการข้อมูลขนาดใหญ่ เช่น การจัดเก็บ การกลั่นกรองเลือกข้อมูล การแปลงลักษณะของข้อมูลให้เหมาะสมกับการนำไปวิเคราะห์ รวมถึงการนำไปสร้างเป็นซอฟต์แวร์เพื่อนำข้อมูลที่เป็นประโยชน์ไปใช้เพิ่มประสิทธิภาพงานในสถานการจริงกับองค์กร เพราะคอมพิวเตอร์มีสมรรถนะในด้าน ความเร็ว ความจุ ความแม่นยำ ที่มีประสิทธิภาพมากกว่าการให้มนุษย์ลงมือปฏิบัติเอง โดยขั้นตอนการใช้ศาสตร์ด้านวิทยาการข้อมูลในการสกัดหาข้อมูลที่เป็นประโยชน์จะกล่าวถึงในหัวข้อถัดไป

การสกัดข้อมูลที่เป็นประโยชน์จากข้อมูลขนาดใหญ่ทำให้สามารถเห็นรูปแบบความสัมพันธ์ของข้อมูลที่เป็นประโยชน์ที่ซ้อนอยู่ในข้อมูลขนาดใหญ่ ซึ่งอาจเป็นประโยชน์ต่อกิจกรรมขององค์กร และสามารถนำมาประยุกต์ใช้เพื่อเพิ่มประสิทธิภาพกิจกรรมขององค์กร เช่น การวางแผนการตลาดที่มีประสิทธิภาพ สร้างโอกาสในการสร้างผลกำไร ทำให้เราเกิดการได้เปรียบเหนือคู่แข่งทางการตลาดได้ การวางแผนการผลิตที่สามารถพยากรณ์ล่วงหน้าในการผลิตสินค้า ซึ่งทำให้องค์กรสามรถจัดการต้นทุนการผลิตได้ เป็นต้น

ดังนั้น วัตถุประสงค์ของเอกสารฉบับนี้ เพื่อให้เกิดความเข้าใจถึงวิธีการค้นหาข้อมูลที่เป็นประโยชน์จากข้อมูลขนาดใหญ่เพื่อนำไปเพิ่มประสิทธิภาพกับองค์กร และสาเหตุการเกิดข้อมูลขนาดใหญ่ การจัดการกับข้อมูลขนาดใหญ่เพื่อค้นหารูปแบบความสัมพันธ์ของข้อมูลที่มีความสำคัญที่ซ้อนอยู่ในคลังข้อมูลขนาดใหญ่ตามความต้องการของผู้ใช้ ตามวัตถุประสงค์การหากฎความสัมพันธ์ด้วย Apriori algorithm

2. ข้อมูลขนาดใหญ่ (Big Data)

ข้อมูล [2] คือ ข้อเท็จจริงหรือสิ่งที่ถือหรือยอมรับว่าเป็นข้อเท็จจริง สำหรับใช้เป็นหลักอนุมานหาความจริงหรือการคำนวณ ซึ่งข่าวสาร ข้อเท็จริงนี้อาจอยู่ในรูปของตัวเลข ภาพ เสียง สัญลักษณ์ต่างๆ ที่มีความหมายเฉพาะตัว และยังไม่มีการประมวล ไม่เกี่ยวกับการนำไปใช้ได้อย่างมีประสิทธิภาพ [3] ในยุคดิจิทัลนี้สิ่งที่จำเป็นและขาดไม่ได้ก็คือข้อมูล เพราะมนุษย์หันมาใช้เทคโนโลยีในการติดต่อสื่อสาร จากสถิติปี 2560 นั้นคนทั่วโลกใช้โซเชียลมีเดีย (Social Media) มากถึง 2.7 พันล้านคน คิดเป็นร้อยละ 37 ของคนทั่วโลก โดยจากคนที่ใช้โซเชียลมีเดีย มีถึงร้อยละ 91 ที่ใช้งานผ่านอุปกรณ์เคลื่อนที่ขนาดเล็ก [4] จากข้อมูลนี้ถือว่าเป็นปัจจัยหนึ่งที่ทำให้เกิดข้อมูลขนาดมหาศาล รวมถึงอุปกรณ์หน่วยความจำที่มีขนาดความจุมากขึ้นซึ่งสวนทางกับราคาที่ถูกลงทำให้เราสามารถจัดเก็บข้อมูลได้อย่างมหาศาล จนเกิดเป็นข้อมูลขนาดใหญ่ ซึ่งเป็นสิ่งที่กำลังได้รับความสนใจเป็นอย่างมากในปัจจุบันนี้ เพราะมีหลายองค์กรที่ใช้ข้อมูลในการขับเคลื่อนหรือใช้พึ่งพิงในการช่วยตัดสินใจในเรื่องต่างๆ รวมถึงการนำเข้ามาใช้ในการพยากรณ์เหตุการณ์ล่วงหน้าของกิจกรรมในองค์กร

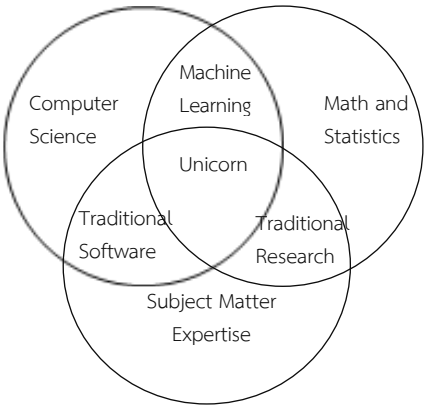
ข้อมูลขนาดใหญ่ หรือ อีกชื่อคือ “บิ๊กดาต้า” ที่จริงแล้วก็มีลักษณะเหมือนข้อมูลขนาดเล็กทั่วๆ ไปแต่มีขนาด หรือปริมาณของข้อมูลขนาดใหญ่กว่าหลายสิบหลายร้อยเท่านั้นเอง โดยสามารถอธิบายด้วยลักษณะ 4 ประการ (4V) [5],[6],[7] ดังต่อไปนี้

- ปริมาณ (Volume) ที่มีขนาดโตมากอยู่ในระดับเทอร์ร่าไบท์ (Terra byte) ไปถึงระดับเพต้าไบท์ (Petabyte) ดังนั้นจึงต้องมีเครื่องมือที่เข้ามาช่วยในการจัดการข้อมูลขนาดใหญ่
- ความเร็ว (Velocity) ของข้อมูลที่มีการเคลื่อนไหวในยุกดิจิทัลที่ในอดีตอยู่ในรูปแบบแบทช์ (Batch) จนกลายมาเป็นแบบทันทีทันใด (Real-Time)
- ความหลากหลาย (Variety) ของข้อมูลที่อาจเป็นข้อมูลภาพ เสียง ข้อความที่อยู่ในรูปแบบที่มีโครงสร้าง แบบไม่มีโครงสร้าง หรือแบบกึ่งโครงสร้าง

- คุณภาพของข้อมูล (Veracity) คุณภาพของข้อมูลที่ถูกบันทึกไว้อาจจะมีความแตกต่างกันมากซึ่งส่งผลต่อการวิเคราะห์ที่ถูกต้องได้ [8] ข้อมูลจาก 4วี (4V) ข้างต้นถือว่าเป็นลักษณะของข้อมูลขนาดใหญ่ ซึ่งข้อมูลขนาดใหญ่ที่เกิดขึ้นในองค์กรเกิดมาจาก 5 แหล่ง [9] ได้แก่ ข้อมูลองค์กร ข้อมูลการทำธุรกรรม สื่อสังคม ข้อมูลสาธารณะ และ ข้อมูลเซนเซอร์ผนวกเข้ากับความสำคัญของข้อมูลที่มีผลต่อการดำเนินธุรกิจ หรือกิจกรรมต่างๆ ที่ส่งเสริมประสิทธิภาพขององค์กรได้นั้น จึงมีนักวิจัย นักวิชาการให้ความสนใจและได้คิดค้นกระบวนการเข้ามาสกัดหรือค้นหารูปแบบความสัมพันธ์ของข้อมูลที่มีความสำคัญที่ซ้อนอยู่ภายในข้อมูลอันมหึมาบนโลกโซเชียล หรือที่ถูกเรียกกันว่ายุคดิจิทัล และในปัจจุบันนี้รัฐบาลของประเทศไทยเราก็มองเห็นถึงความสำคัญและได้ปรับแนวทางของประเทศให้เป็นยุคสมัยที่ 4 ที่เราได้ยินกันในนิยามไทยแลนด์ 4.0 [10] ซึ่งก็คือยุคที่ต้องใช้เครื่องมือทางเทคโนโลยีคอมพิวเตอร์เข้ามาช่วยในการจัดการกิจกรรมต่างๆ ภายในประเทศด้วยการใช้ข้อมูลเป็นตัวขับเคลื่อนไม่ว่าจะเป็นทางด้านการศึกษา เศรษฐกิจ การเมือง ธุรกิจ และด้านต่างๆ ที่มีผลต่อการส่งเสริมกิจกรรม การพัฒนาประเทศ และก้าวทันอารยประเทศที่พัฒนาแล้ว แต่สิ่งที่เป็นปัญหาตามมากับข้อมูลที่มีขนาดมหึมานี้คือการเข้าไปสกัดหาข้อมูลที่มีประโยชน์และต้องได้ข้อมูลที่มีประโยชน์ที่แท้จริง เพื่อจะได้นำมาใช้ประโยชน์ได้อย่างมีประสิทธิภาพ ถือว่าเป็นความท้าทายมิใช่น้อย จึงได้เกิดศาสตร์ทางด้านวิทยาการข้อมูล (Data Science) ซึ่งเป็นศาสตร์ที่มีกระบวนการ ขั้นตอนที่เป็นวิทยาศาสตร์เข้าจัดการกับข้อมูลมหาศาล ที่จะกล่าวถึงในหัวข้อถัดจากนี้

3. วิทยาการข้อมูล (Data Science)

ในศตวรรษที่ 21 โลกธุรกิจต่างๆ ได้หันมาให้ความสำคัญกับข้อมูลขนาดใหญ่เพิ่มมมากขึ้น [11] เพราะในข้อมูลขนาดใหญ่นั้นจะมีรูปแบบความสัมพันธ์ของข้อมูลที่มีประโยชน์ต่อการนำมาต่อยอด นำมาตัดสินใจ หรือแม้แต่นำมาช่วยปรับแผนกลยุทธ์ขององค์กรได้อย่างมีประสิทธิภาพ และสร้างความได้เปรียบกับคู่แข่ง ดังตัวอย่างที่เราสามารถเห็นได้ เช่น การซื้อของออนไลน์เห็นได้ว่าระบบสามารถนำเสนอสินค้าที่ตรงตามความสนใจของลูกค้ามานำเสนอเพื่อให้เกิดโอกาสการซื้อขายสูงขึ้น ที่เป็นเช่นนี้ได้เพราะมีกระบวนการทางวิทยาศาสตร์ที่เข้ามาช่วยในการกลั่นกรองข้อมูลขนาดใหญ่ที่มีอยู่ เพื่อหารูปแบบของข้อมูลที่มีประโยชน์ โดยกระบวนการทางวิทยาศาสตร์ที่เป็นการผนวกกันขององค์ความรู้ด้านสถิติ คณิตศาสตร์ การเขียนโปรแกรมทางคอมพิวเตอร์ และเทคโนโลยี เพื่อตั้งสมมติฐาน ทดลอง และหาผลลัพธ์จากข้อมูลขนาดใหญ่ ซึ่งถูกเรียกว่า วิทยาการข้อมูล หรือ Data Science สารมรถสรุปนิยามได้ว่า เป็นศาสตร์ที่ทำการศึกษาการสกัดความรู้จากข้อมูล [12] ซึ่งได้แสดงเป็นแผนภาพความสัมพันธ์ดังภาพที่ 1 [13]



ภาพที่ 1 การบูรณาการของวิทยาการข้อมูล

4. กระบวนการสกัดข้อมูลที่เป็นประโยชน์จากข้อมูลขนาดใหญ่

เมื่อมีข้อมูลขนาดใหญ่แล้ว สิ่งที่จะต้องดำเนินการต่อไปก็คือการเข้าไปค้นหาข้อมูลที่เป็นประโยชน์ หรือรูปแบบความสัมพันธ์ของข้อมูลที่จะมีประโยชน์ต่อการไปปรับใช้กับองค์กร ก็คือกระบวนการวิเคราะห์ข้อมูลขนาดใหญ่ด้วยกระบวนการทำงานที่เรียกว่า “Cross-Industry Standard Process for Data Mining” หรือเรียกย่อ ว่า “CRISP-DM” [14],[15],[16] ถือได้ว่าเป็นส่วนหนึ่งของ วิทยาศาสตร์ข้อมูล เพราะกระบวนการนี้ใช้เป็นการตั้งสมมติฐาน ทดลอง และหาผลลัพธ์จากข้อมูลขนาดใหญ่ โดยมีขั้นตอน 6 ขั้นตอนด้วยกัน ดังแสดงในภาพที่ 2 มีรายละเอียดดังต่อไปนี้

4.1. การทำความเข้าใจธุรกิจ (Business Understanding) เป็นขั้นตอนแรกของการวิเคราะห์ข้อมูลขนาดใหญ่ คือการเน้นความเข้าใจกับปัญหา หรือโอกาสเชิงธุรกิจ แล้วระบุเป้าหมายที่ต้องการได้จากการวิเคราะห์ข้อมูลขนาดใหญ่ เช่น ทำอย่างไรให้เพิ่มยอดขาย

4.2. การทำความเข้าใจกับข้อมูล (Data Understanding) เป็นการทำการตรวจสอบความถูกต้องและการเลือกว่าจะใช้ข้อมูลทั้งหมดหรือบางส่วนจากข้อมูลที่ได้รวมรวมไว้เพื่อจะนำมาทำการวิเคราะห์

4.3. การจัดเตรียมข้อมูล (Data Preparation) เป็นการแปลงข้อมูลจากขั้นที่แล้วให้อยู่ในรูปแบบที่สามารถนำเข้าไปใช้กับแบบจำรองในขั้นถัดไป โดยการแปลงในขั้นตอนนี้เป็นการทำให้ข้อมูลสะอาด เช่น ข้อมูลที่เก็บมาอาจจะขาดหายหรือไม่สมบูรณ์ หรือข้อมูลอยู่ช่วงที่แตกต่างกัน จะต้อง

ปรับแก้เพิ่มเติมให้ข้อมูลสมบูรณ์ โดยใช้หลักการทางสถิติคณิตศาสตร์ เข้ามาเพิ่มเติมข้อมูลให้สมบูรณ์ ขั้นตอนนี้จะกินเวลาจากทุกขั้นตอนถึง 80% เพราะถ้าข้อมูลไม่ถูกต้องและไม่สมบูรณ์ก็จะทำให้นำไปสร้างแบบจำลองแล้วได้ผลลัพธ์ออกมาไม่ถูกต้อง

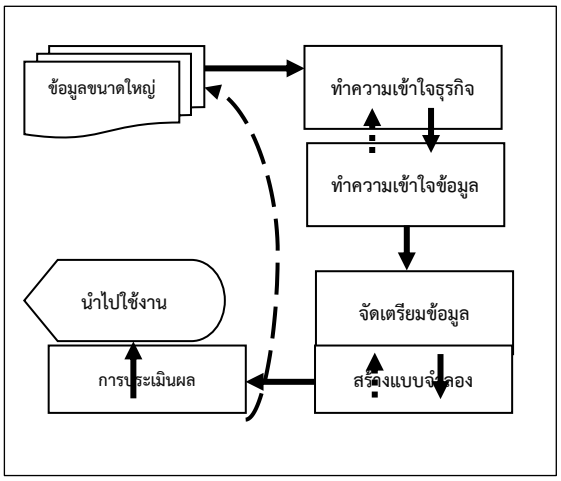
4.4. การสร้างแบบจำลอง (Modeling) เป็นขั้นตอนการวิเคราะห์ข้อมูลตามวัตถุประสงค์ 3 วัตถุประสงค์ คือ

- การหากฎความสัมพันธ์ (Association Rules)
- การแบ่งกลุ่มข้อมูล (Clustering)
- การจำแนกประเภทข้อมูล (Classification)

ซึ่งจะได้อธิบายเพิ่มเติมในหัวข้อถัดไป

4.5. การประเมินผล (Evaluation) หลังจากได้ผลลัพธ์จากการวิเคราะห์จากขั้นที่ 4 ก่อนจะนำผลวิเคราะห์ไปใช้จะต้องทำการวัดประสิทธิภาพของผลการวิเคราะห์นั้นเสียก่อนว่าตรงตามวัตถุประสงค์ และมีความน่าเชื่อถือของผลวิเคราะห์นั้นมากน้อยเพียงใด

4.6. การปรับใช้ (Deployment) ในการวิเคราะห์ข้อมูลขนาดใหญ่ไม่ได้จบอยู่แค่เพียงขั้นตอนประเมินผลเท่านั้น แม้ว่าผลลัพธ์ที่ได้จะแสดงถึงรูปแบบความสัมพันธ์ของข้อมูลที่มีประโยชน์ เราจะต้องนำเอาองค์ความรู้นี้ไปใช้ประโยชน์จริงในองค์กร เช่น การทำรายงานให้กับผู้บริหารดูเข้าใจง่าย เพื่อจะนำไปปรับกลยุทธ์ เป็นต้น

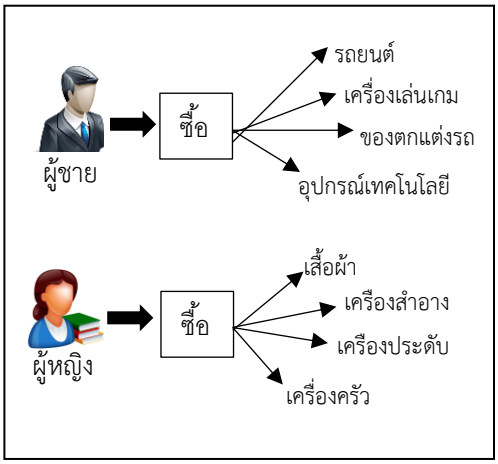


ภาพที่ 2 แสดงขั้นตอนในกระบวนการ CRISP-DM

5. วัตถุประสงค์การสกัดข้อมูลที่เป็นประโยชน์จากข้อมูลขนาดใหญ่

ในขั้นตอนสร้างแบบจำลอง (Modeling) ของกระบวนการวิเคราะห์ข้อมูลขนาดใหญ่ (CRISP-DM) [17] เป็นขั้นตอนที่ให้ผลลัพธ์การสกัดข้อมูลที่สำคัญให้กับผู้วิเคราะห์ข้อมูลขนาดใหญ่ตามวัตถุประสงค์ที่ผู้วิเคราะห์ต้องการ โดยมีด้วยกัน 3 วัตถุประสงค์ ดังต่อไปนี้

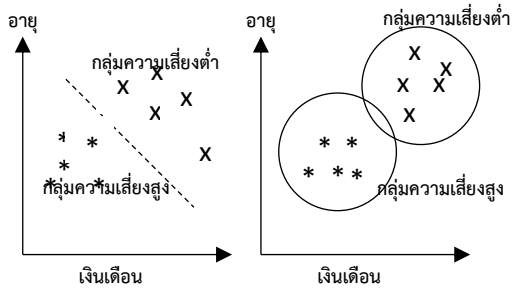
5.1 การหากฎความสัมพันธ์ (Association Rules) ของข้อมูล คือ กระบวนการกฎการเรียนรู้ด้วยเครื่อง สำหรับการค้นพบความสัมพันธ์ที่น่าสนใจระหว่างตัวแปรในฐานข้อมูลขนาดใหญ่ [18] ซึ่งเป็นการอธิบายการวิเคราะห์และค้นหากฎที่มีความน่าเชื่อถือมากที่สุดจากกฎทั้งหมดที่หาได้จากข้อมูลขนาดใหญ่ โดยอัลกอริธึมที่นิยมใช้ในการหากกฎความสัมพันธ์คือ Apriori Algorithm [19] ผลที่ได้จากวัตถุประสงค์นี้คือความสัมพันธ์กันของข้อมูลดังภาพที่ 3



ภาพที่ 3 ความสำพันธ์ระหว่างการเลือกรายการสินค้า

5.2 การจัดกลุ่มข้อมูล (Clustering) ถือว่าเป็นปัญหาการเรียนรู้ที่ไม่มีผู้สอน คือการจัดกลุ่มข้อมูลจะไม่ได้ถูกจัดกลุ่มหรือกำหนดกลุ่มข้อมูลไว้ล่วงหน้า แต่จะเป็นการใช้ความคล้ายกันของวัตถุเป็นเกณฑ์ในการจัดกลุ่มข้อมูลนั้นๆ ให้เป็นกลุ่มเดียวกัน [20] โดยข้อมูลที่อยู่ต่างกลุ่มกันก็จะไม่มีความคล้ายกัน

5.3 การจำแนกประเภทข้อมูล (Classification) คือ กระบวนการสร้างโมเดลจัดการข้อมูลให้อยู่ในกลุ่มที่กำหนดมาให้จากกลุ่มตัวอย่างข้อมูลที่เรียกว่าข้อมูลสอนระบบ [21] การจำแนกข้อมูลสามารถพิจารณากลุ่มให้กับข้อมูลที่ยังไม่ได้แบ่งกลุ่มในอนาคตได้ ซึ่งความต่างของการจำแนกกลุ่มข้อมูลและการจัดกลุ่มข้อมูล คือ การจำแนกกลุ่มจะเหมาะกับปัญหาที่ต้องมีการสอนระบบการจำแนก ซึ่งจะต้องมีการกำหนดกลุ่มของข้อมูลไว้ก่อนแล้วนำข้อมูลที่มีความคล้ายกลุ่มที่ได้กำหนดไว้ก่อนหน้าเข้าไปในกลุ่มนั้นๆ ส่วนการจัดกลุ่มจะไม่สามารถทราบกลุ่มล่วงหน้าหรือไม่ได้มีการจัดกลุ่มไว้ล่วงหน้าแต่จะจัดกลุ่มตามความคล้ายกันของข้อมูลทั้งหมดที่ได้ทำการวิเคราะห์ โดยภาพที่ 4 ได้แสดงถึงความแตกต่างระหว่างการจัดกลุ่มและการจำแนกกลุ่มข้อมูล



ภาพที่ 4 การจำแนกกลุ่มความเสี่ยงสำหรับการปล่อยเงินกู้ให้กับพนักงานเปรียบเทียบกับการจัดกลุ่ม

6. การประยุกต์ใช้งานการสกัดข้อมูลที่เป็นประโยชน์จากข้อมูลขนาดใหญ่ตามวัตถุประสงค์กฎความสัมพันธ์ด้วย Apriori algorithm

การหากฎความสัมพันธ์ในเอกสารนี้จะขอนำเสนอแค่ Apriori algorithm [19][21] เนื่องจากเป็นอัลกอริธึมที่เข้าใจง่าย โดยมีอยู่ 2 ขั้นตอน ดังนี้

ขั้นที่หนึ่ง การหา Frequent item set คือการหารูปแบบของข้อมูลที่เกิดขึ้นร่วมกันบ่อยๆ ในฐานข้อมูล โดยความถี่ที่เกิดขึ้นร่วมกันนั้นจะต้องมีค่ามากกว่าค่า minimum support [22] โดยสามารถคำนวณตามสมการที่ 1

$$supp(x) = \frac{|\{t \in T; X \subseteq t\}|}{T} \qquad (1)$$

ในบทความนี้ขอใช้ตัวอย่างข้อมูลดังแสดงในตารางที่ 1 ซึ่งเป็นรายการซื้อขายสินค้าจากฐานข้อมูล

ตารางที่1 แสดงรายการซื้อสินค้าที่จะใช้ในการหากฎความสัมพันธ์

เลขที่ใบเสร็จ	รายการสินค้า
001	ทุเรียน ไข่ไก่ นมผง
002	กาแฟ นมผง ไข่ไก่
003	ทุเรียน ไข่ไก่ นมผง กาแฟ
004	กาแฟ ไข่ไก่

ทำการหาค่า minimum support แต่ละสินค้า (item) ซึ่งได้ผลตามตารางที่ 2 กำหนดค่า minimum support เท่ากับ 70% ซึ่งค่านี้สามารถกำหนดเองได้ ถ้ากำหนดค่าสูงแสดงว่าความกี่ในการเกิดจะต้องมีความถี่มากๆ ตาม

ตารางที่ 2 แสดงการคำนวณค่า support ของสินค้า

สินค้า	เลขที่ใบเสร็จ				Support
	001	002	003	004	
ทุเรียน	1	0	1	0	2/4=50%
ไข่ไก่	1	1	1	1	4/4=100%
นมผง	1	1	1	0	3/4=75%
กาแฟ	0	1	1	1	3/4=75%

จากตารางที่ 2 จะเห็นว่าสินค้าบางชนิดมีค่า minimum support ไม่ถึง 70% ตามที่กำหนดไว้ให้ทำการตัดทิ้งและไม่นำสินค้าชนิดนั้นไปดำเนินการในขั้นถัดๆ ไปเนื่องจากความถี่ที่เกิดขึ้นน้อยกว่าที่กำหนด หรือถ้าพูดอีกนัยหนึ่งการซื้อสินค้าชนิดนี้ที่ยังน้อยอยู่ ซึ่งก็คือ “ทุเรียน”

จากนั้นทำการเพิ่มความยาวจำนวน สินค้า จากหนึ่งชนิดเป็นสองชนิดต่อหนึ่งเลขที่ใบเสร็จ โดยปราศจาก สินค้าที่มีค่า minimum support ต่ำกว่าที่กำหนด และทำการหา minimum support ใหม่อีกรอบ ซึ่งได้ผลตามตารางที่ 3 และทำการตัดชุดสินค้าที่มีค่า minimum support ที่ต่ำกว่า 70% ดังนั้นสินค้าที่อยู่ในชุดสินค้านั้นจะไม่ถูกทำไปใช้ในขั้นที่มีชุดสินค้าเพิ่มขึ้นด้วย และทำการเพิ่มความยาวของชุดสินค้าเพื่อหาค่า minimum support และตัดทิ้งสำหรับสินค้าที่มีค่า minimum support ต่ำกว่ากำหนด ทำแบบนี้จนกว่าจะไม่สามารถเพิ่มความยามชุดสินค้าได้ ซึ่งในบทความนี้จะเหลือเพียง 2 ชุดสินค้าดังแสดงในตารางที่ 4 และไม่สามารถเพิ่มชุดสินค้าเป็นความยาวสามได้

ตารางที่ 3 แสดงค่า support ของชุดสินค้าความยาวสอง

สินค้า	เลขที่ใบเสร็จ				Support
	001	002	003	004	
{ไข่ไก่,นมผง}	1	1	1	0	3/4=75%
{ไข่ไก่,กาแฟ}	0	1	1	1	3/4=75%
{นมผง,กาแฟ}	0	1	1	0	2/4=50%

ตารางที่ 4 แสดงชุดสินค้าที่เหลือที่มีค่าเท่ากับหรือมากกว่าค่า minimum support

สินค้า	เลขที่ใบเสร็จ				Support
	001	002	003	004	
{ไข่ไก่,นมผง}	1	1	1	0	3/4=75%
{ไข่ไก่,กาแฟ}	0	1	1	1	3/4=75%

แล้วนำเอาชุดสินค้าที่มีค่า minimum support ในแต่ละรอบขนาดความยาวของชุดสินค้ามาสร้างตารางสรุปเป็น Frequent item sets ตามตารางที่ 5

ตารางที่ 5 แสดง Frequent item set ทั้งหมดที่หาได้

Frequent item sets	Support	Size
{ไข่ไก่}	100%	1
{นมผง}	75%	1
{กาแฟ}	75%	1
{ไข่ไก่,นมผง}	75%	2
{ไข่ไก่,กาแฟ}	75%	2

ขั้นที่สอง ขั้นตอนการสร้างกฎความสัมพันธ์ซึ่งจะเกิดขึ้นหลังจากทำขั้นที่หนึ่ง การหา Frequent item set โดยนำ

ผลสรุปจากขั้นที่หนึ่งนั้นมาพิจารณาด้วยการวัดประสิทธิภาพของกฎซึ่งในบทความนี้จะใช้ค่า confidence และค่า lift ซึ่งเป็นไปตามสมการที่ 2 และ 3 ตามลำดับ

ค่า confidence [22][23] เป็นการแสดงความเชื่อมันของกฎความสัมพันธ์เมื่อ x เกิดขึ้นแล้ว y จะเกิดขึ้นตามคิดเป็นกี่เปอร์เซ็นต์ โดยสามารถคำนวณได้ตามสมการที่ 2

$$conf(X \Rightarrow Y) = \frac{supp(x \cup y)}{supp(x)} \qquad (2)$$

ค่า lift [22][23] คือค่าที่บ่งบอกว่าการเกิดรูปแบบ x และ y มีความสัมพันธ์กันแค่ไหน ถ้าค่า lift เป็น 1 แสดงว่า x และ y ไม่ขึ้นต่อกัน ซึ่งสามารถคำนวณได้จากสมการที่ 3

$$lift(x \Rightarrow y) = \frac{supp(x \cup y)}{supp(x) \times supp(y)} \qquad (3)$$

หลังจากนำข้อมูลจากขั้นที่หนึ่งตามตารางที่ 5 มาทำการหาค่า confidence และค่า lift แล้ว ได้ผลตามตารางที่ 6

ตารางที่ 6 แสดงกฎความสัมพันธ์ทั้งหมดที่สร้างได้พร้อมทั้งค่า confidence และ lift

ลำดับ	Frequent itemset	conf	lift
1	{ไข่ไก่ => นมผง}	50%	0.66
2	{ไข่ไก่=>กาแฟ}	25%	0.33
3	{กาแฟ => ไข่ไก่}	50%	0.66
4	{นมผง => ไข่ไก่}	25%	0.33
5	{กาแฟ => นมผง}	25%	0.33
6	{นมผง => กาแฟ}	25%	0.33

7. สรุปผลการศึกษา

ปัจจุบันเกือบทุกองค์กรหันมาใช้เทคโนโลยีสารสนเทศเข้ามาช่วยในการเก็บรวบรวมข้อมูลและด้วยความสามารถของเทคโนโลยีที่บรรจุข้อมูลขนาดใหญ่ไว้ได้จึงส่งผลให้ข้อมูลมีขนาดใหญ่ที่ปะปนไปด้วยข้อมูลที่มีประโยชน์ซึ่งข้อมูลนี้จะสามารถนำไปสร้างประโยชน์หรือเพิ่มประสิทธิภาพขององค์กรได้และอีกส่วนก็เป็นข้อมูลที่ไม่สามารถนำไปใช้ประโยชน์ได้ เมื่อพิจารณาแล้วน่าจะเป็นอุปสรรค์กับหลายๆ องค์กรแต่ด้วยความฉลาดของมนุษย์เราที่มีการพัฒนาแบบไม่หยุดยั้งจึงทำให้เรามีกระบวนการสกัดข้อมูลที่เป็นประโยชน์จากข้อมูลขนาดใหญ่ด้วยกระบวนการ “Cross-Industry Standard Process for Data Mining” ดังที่ได้กว่าวไว้ในหัวข้อที่ 4 และในขั้นตอนการสร้างแบบจำลอง (Modeling) ของกระบวนการ ซึ่งมีนักวิจัยได้คิดค้นอัลกอริธึมอีกหลายร้อยแบบมาช่วยในการสกัดหาข้อมูลที่เป็นประโยชน์เพื่อนำไปใช้เพิ่มประสิทธิภาพขององค์กร ในบทความนี้ก็ได้นำเสนอ Apirori algorithm ที่สามารถนำไปประยุกต์ใช้กับการสกัดหาข้อมูลที่เป็นประโยชน์ได้จากข้อมูลขององค์กรที่ได้จัดเก็บไว้

เอกสารอ้างอิง

[1] รังสรรค์ ประเสริฐศรี. (2548). พฤติกรรมองค์กร กรุงเทพ ฯ : ธรรมสาร

[2] ราชบัณฑิตยสถาน. (2556). พจนานุกรม ฉบับราชบัณฑิตยสถาน พ.ศ. ๒๕๕๔ เฉลิมพระเกียรติพระบาทสมเด็จพระเจ้าอยู่หัว เนื่องในโอกาสพระราชพิธีมหามงคลเฉลิมพระชนมพรรษา ๗ รอบ ๕ ธันวาคม ๒๕๕๔. กรุงเทพฯ : ราชบัณฑิตยสถาน

[3] ไพโรจน์ คชชา. (2542). คอมพิวเตอร์และเทคโนโลยีสารสนเทศสำหรับผู้บริหาร. กรุงเทพฯ: เซนเตอร์ดิสคัฟเวอรี่.

[4] WeAre Social and Hootsuite. (2017) Digital in 2017 global overview.[4 Mar 2018].https://www.slideshare.net/wearesocialsg/digital-in-2017-global-overview

[5] A. Adrian. “Big Data Challenges”. (2013). American University

[6] Laney, Doug. (2001). "3D data management: Controlling data volume, velocity and variety". META Group Research Note. 6 (70).

[7] Hilbert, Martin. (2015) "Big Data for Development: A Review of Promises and Chal-lenges. Development PolicyReview". Martinhilbert.net.

[8] Spotless Data. (2017). BIG DATA'S FOURTH V. [Online]. https://spotlessdata.com/blog/big-datas-fourth-v

[9] Siddharth Singh,Tuba Firdaus and Dr. A.K. Sharma.(2015). Survey on Big Data Using Data Mining. International Journal of Engineering Development and Research. Volume 3, Issue 4.P135-143

[10] กองบริหารงานวิจัยและประกันคุณภาพการศึกษา มหาวิทยาลัยพะเยา.(2559). โมเดลขับเคลื่อนประเทสไทยสู่ความมั่งคั่ง มั่นคง และยั่งยืน.[Online]. http://www.libarts.up.ac.th/v2/img/Thailand- 4.0.pdf

[11] Thomas H. DavenportD.J. Patil.(2012).Data Scientist: The Sexiest Job of the 21st Century.[Online].https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st- century

[12] [Dhar , V. (2013). "Data science andprediction". Communications of the ACM.P.64.

[13] [datanami. (2016). As Data Science Evolves, It's Taking Statistics with It. [Online]. https://www.datanami.com

[14] [Apipoj Piasak. (2017). [Online].What is Data Science.https://blog.derlivery.com

[15] C. Shearer, "The CRISP-DM model: The new blueprint for data mining". Journal of Data Warehousing, 5(4), 13–22, 2000

[16] เอกสิทธิ์ พัชรวงศ์ศักดา.(2017). Practical Data Mining with Rapidminer Studio 7. กรุงเทพมหานคร : เอเชีย ดิจิตอลการพิมพ์

[17] Data Mining Trend. (2014).[Online]. กระบวนการวิเคราะห์ข้อมูลด้วย CRISP-DM และ ตัวอย่างการประยุกต์ใช้ทางด้านการศึกษา. http://dataminingtrend.com/2014/data-mining-techniques/crisp-dm-example/

[18] G. Piatetsky-Shapiro (1991), Discovery, Analysis, and Presentation of Strong Rules",pp. 229-248.

[19] Rakesh Agrawal, Tomasz Imielinski and Arun Swami .(1993). Mining Association Rules between Sets of Items in Large Databases. Proceedings of the 1993 ACM SIGMOD Conference Washington DC, USA

[20] Jia Li. (2013) "Data Mining - Clustering by MixtureModels".[Online] http://www.stat.psu.edu/~jiali/course/stat597e/notes/mix.pdf

[21] Data Mining Trend. (2014). "ขั้นตอนการหากฎความสัมพันธ์ (Association Rules)". [Online]. http://dataminingtrend.com/2014/association-rules/

[22] Hahsler, Michael (2005). "Introduction to arules – A computational environment for mining association rules and frequent item sets" . Journal of Statistical Software.

[23] Hipp, J.; Güntzer, U.; Nakhaeizadeh, G. (2000). "Algorithms for association rule mining ageneral survey and comparison". ACM SIGKDD Explorations Newsletter. 2: 58